DTU Library

Reliability of adaptive multivariate software sensors for sewer water quality monitoring

Vezzaro, Luca; Mikkelsen, Peter Steen

Published in:
Proceedings of 10th International Urban Drainage Modelling Conference

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Reliability of adaptive multivariate software sensors for sewer water quality monitoring

Luca Vezzaro[1], Peter Steen Mikkelsen[1]

[1]Department of Environmental Engineering, Technical University of Denmark, Building 113, Miljoevej, 2800 Kgs. Lyngby, Denmark (Email: *luve@env.dtu.dk; psmi@env.dtu.dk*)

Abstract
This study investigates the use of a multivariate approach, based on Principal Component Analysis PCA), as software sensor for fault detection and reconstruction of missing measurements in on-line monitoring of sewer water quality. The analysis was carried out on a 16-months dataset of five commonly available on-line measurements (flow, turbidity, ammonia, conductivity and temperature). The results confirmed the great performance of PCA (up to 10 weeks after parameter estimation) when estimating a measurement from the combination of the remaining four variables, a useful feature in data validation. However, the study also showed a dramatic drop in predictive capability of the software sensor when used for reconstructing missing values, with performance quickly deteriorating after 1 week since parameter estimation. The software sensor provided better results when used to estimate pollutants mainly originated from wastewater sources (such as ammonia) than when used for pollutants affected by several processes (such as TSS). Overall, this study provides a first insight in the application of multivariate methods for software sensors, highlighting drawback and potential development areas. A combination of (i) advanced methods for on-line data validation, (ii) frequent parameter estimation, and (iii) automatic method for classification of dry/wet periods may provide the needed background for a successful application of these software sensors.

Keywords
On-line water quality monitoring; Principal component analysis; Software sensors; data quality control;

INTRODUCTION

On-line sensors for monitoring water quality parameters in sewers are getting increasing attention (Campisano et al., 2013). These devices can provide high-temporal resolution information on various water quality variables (e.g. turbidity, conductivity, NH_4), which can subsequently be used to analyse and to elaborate statistics on the pollution loads discharged from the urban drainage systems over long time periods (see for example the dataset collected by Metadier and Bertrand-Krajewski, 2012). Also, the on-line measurements from these sensors can be integrated within water-quality based Real Time Control (RTC) strategies, as in the example presented by Lacour and Schuetze (2011). These RTC applications require a constant flow of information, i.e. fall-back strategies need to be available in case of sensors failure. In this context, surrogate measurements can be provided by software sensors, developed for on-line fault detection and data reconstruction (Kadlec et al., 2009).

Physical sensors are often affected by malfunctioning, drifting, erroneous readings, etc., and these issues affect the quality of the collected information. Procedures for Data Quality Control (DQC) of continuous measurements have been developed for on-line monitoring of wastewater treatment plants and they are now extended to sewer monitoring (e.g. Alferes et al., 2013). The most advanced DQC methods utilize multivariate approaches, such as Principal Component Analysis (PCA). PCA can also be employed as software sensor, i.e. missing data from a sensor can be reconstructed by using a combination of other on-line measurements.

However, PCA requires measurements from stationary processes, and thus adaptive approaches have been suggested to ensure their applicability in the highly dynamic field of wastewater systems (Rosen and Lennox, 2001). Also, adaptive approaches require quality-ensured data, i.e. the PCA should be tuned on "good data" (e.g. Rosen and Lennox, 2001). Therefore, advanced multivariate methods, still need to be coupled with simpler univariate DQC approaches, which can both be fully automatic or semi-automatic (i.e. data validation is carried out automatically, but final validation is performed by the operator). Univariate data validation is often based on tests including check of measurement ranges, variation rates, and intervals between maintenance periods.

This study aims at evaluating the performance of multivariate software sensors (using an adaptive PCA) for fault detection and data reconstruction applied to on-line monitoring of sewage quality. On-line wastewater quality data collected during the AMOK project (Sharma et al., 2014) were used in the study, aiming at predicting water quality variables with different characteristics (NH_4, mainly originated by wastewater sources, and Total Suspended Solids – TSS - originated from both waste- and stormwater and affected by various processes along the drainage network). Firstly, it was assessed the potential of predicting a water quality variable based on the combination of existing data. This is a useful feature for fault detection. Subsequently, the performance of the multivariate software sensor for on-line data reconstruction was assessed by using different adaptation techniques, resembling automatic and semi-automatic data validation procedures.

MATERIAL AND METHODS

Physical sensors

The AMOK project (Sharma et al., 2014) aimed at collecting water quality data in a Combined Sewer Overflow, placed at the inlet of the Viby WasteWater Treatment Plant (WWTP), located in Aarhus (Denmark). The upstream catchment is characterized by a mixed structure, with 607 ha drained by combined systems and 925 ha drained by separate systems. Overflows take place when the flow exceeds approximately 1.3 m^3/s, with overflowing water being stored in a detention basin and subsequently pumped to the WWTP after the rain event has ended.

As part of the project, a set of quality sensors were placed *in situ*, i.e. directly in the monitored sewage stream (with a set-up inspired to the one described in Gruber et al., 2005): the data used in this study were collected by these devices from August 2013 to November 2014. The five sensors which have been regarded as the most robust and reliable were used in this study to develop software sensors: flow, pH, conductivity, turbidity (converted into TSS in this study), and NH_4. A Standard Operating Procedure (SOP), similar to the one described in Alferes et al., 2013) was enforced to ensure a good quality of the collected data. The periodical maintenance operations highlighted several issues (sensor drift) which required sensors re-calibration, resulting in data characterized by jumps in the recorded values (see an example in Figure 1). In this study, only simple data validation check (i.e. range checks) were performed, i.e. the data can be affected by e.g. noise (this is quite evident for turbidity measurements).

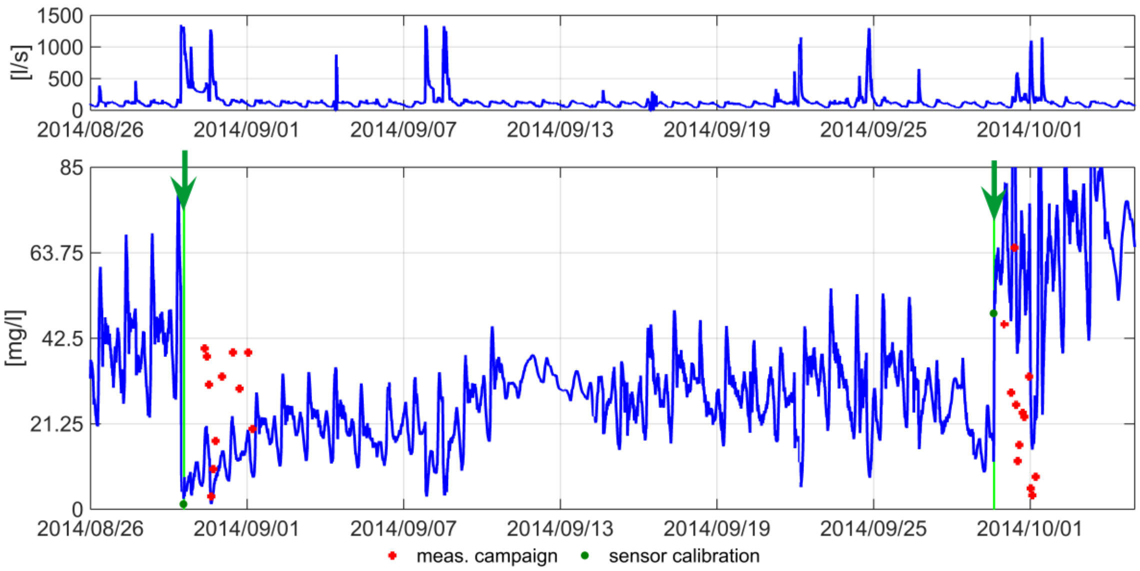

Figure 1. Example of measured flow (above) and on-line NH_4 concentrations (below) between two calibration periods (highlighted in green). Samples collected by autosampler and measured in the laboratory are shown in red.

Multivariate Software sensors

Principal Component Analysis. The approach behind the tested software sensors assumes that the correlations between the measured water quality variables can be mapped by Principal Component Analysis (PCA), whose results are subsequently used to generate predictions (useful as surrogate for missing data). Given that there are N measurements available for calibration for m water quality variable (in this study $m=5$), let's consider the matrix $\boldsymbol{X}$ (N x m), which is built from the standard scores of the available observations $obs_{i,m}$. Each element of $\boldsymbol{X}$ is then calculated by using the average μ_i and standard deviation σ_j^2 of the j-*th* measured variable:

$$x_{i,j} = \frac{(obj_{i,j} - \mu_j)}{\sigma_j^2} \tag{1}$$

The variable to be predicted (in this study NH_4 or TSS) is stored in the m-th column of matrix $\boldsymbol{X}$**,** which can be decomposed as:

$$X = TP^T \tag{2}$$

where the matrix $\boldsymbol{T}$ (N x m) contains the scores (i.e. the observations projected in the principal component space) and $\boldsymbol{P}$ is the matrix containing the loadings. Equation 2 can be also written as:

$$X_{i,j} = \sum_{k=1}^{m} t_{i,k} \cdot p_{j,k} \tag{3}$$

The main assumption behind the software sensor introduced in this study is that the information provided by the m-*1* variables is sufficient to approximate the value of the m-*th* variable, i.e.

$$X_{i,m} \approx X^*_{i,m} = \sum_{k=1}^{m-1} t_{i,k} \cdot p_{m,k} \tag{4}$$

Equation 4 can be useful for fault detection: whenever the estimated values $x^*_{i,m}$ shows a significant deviation from the sensor reading, the latter can be regarded as "doubtful" and not be validated.

PCA-based interpolation: The sensor utilizes a calibration datasets (of size N) to estimate the loadings $p_{i,j}$, stored in the matrix **P**. When new set of available measurements $x'_{1,..,m-1}$ is available, the software sensor reconstructs the *m-th* variable x'_m by the following procedure:

1) Create the projection $\tilde{\mathbf{T}}^{(N+1\times m-1)}$ of the available measurements in the space defined by *m-1* principal components. This step also requires the estimation of the matrix $\tilde{\mathbf{P}}^{(m-1\times m-1)}$, containing the loadings in a component space limited to the available measurements
2) Calculate the projection $\tilde{\mathbf{t}}'$ the new point in the same space
3) Identify the q points in $\tilde{\mathbf{T}}$ which are the closest point to projection $\tilde{\mathbf{t}}'$. In this study the ten closest points were used (i.e. $q=10$).
4) Use the projections of the q points in the m-dimensional space (i.e. their values in $\boldsymbol{T}$) to interpolate the projection $\mathbf{t}'$ of the new measurement. In this study, a simple Inverse-Distance-Weight approach (Li and Heap, 2004) was applied.
5) Calculate the value of the missing variable x'_m by re-transforming the interpolated projection $\mathbf{t}'$ into the original coordinate system by using eq. 4.

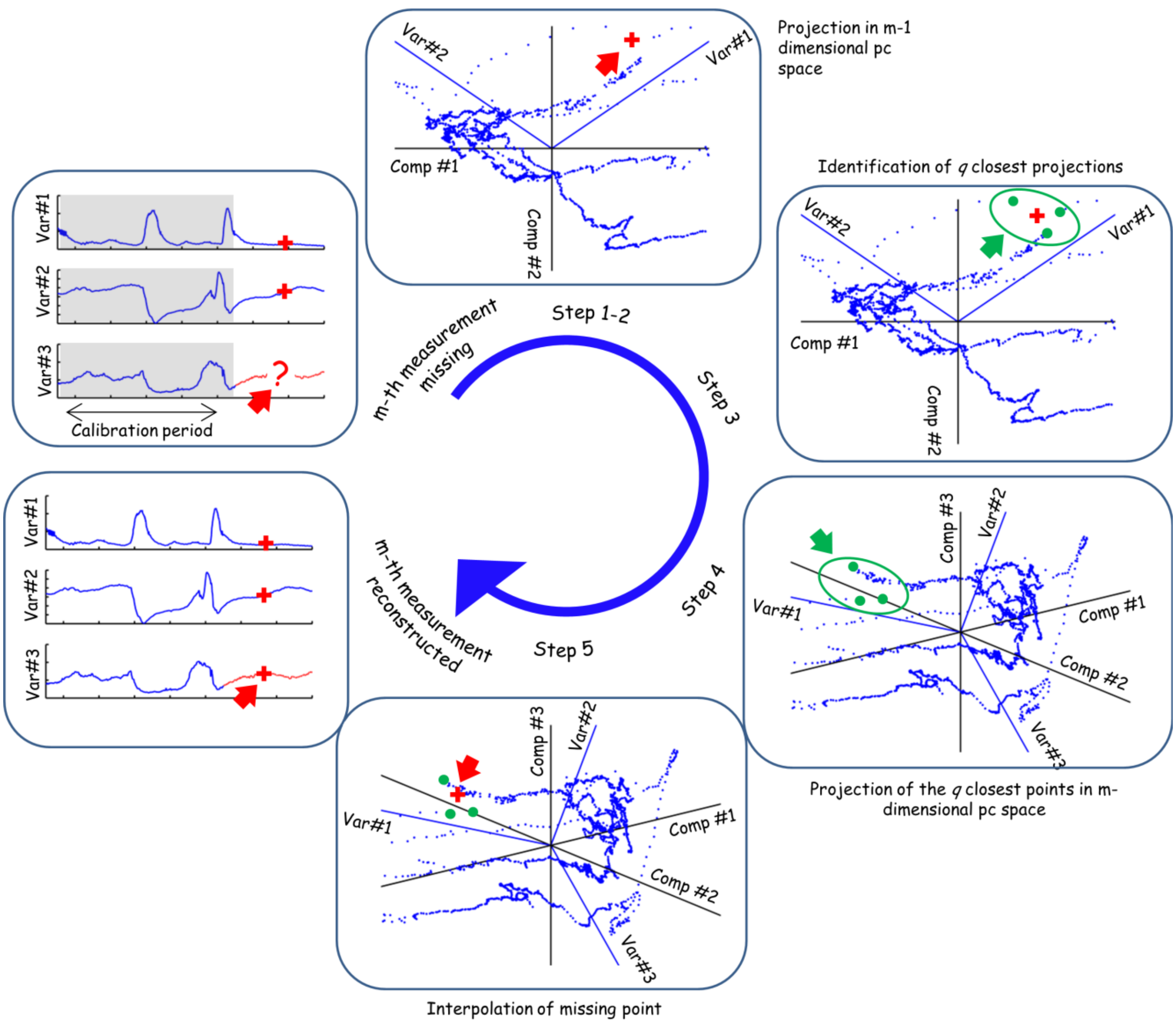

Figure 2. Schematic representation of the steps used by the software sensor to reconstruct a missing measurement (for visualization purpose, only three variables are considered in this figure, i.e. m=3)

Adaptive PCA. The performance of the software sensor described in the previous section are clearly dependent on the information contained in the dataset used for calibration: changes in the monitored system will rapidly lead to a degradation of the sensor's predictions. As discussed in Rosen and Lennox (2001), the limitation of PCA due to non-stationarity of wastewater systems can be overcome by applying *adaptive approaches*. These approaches essentially requires a re-estimation of the model parameters at defined time intervals, i.e. the loadings $p_{i,j}$ should be re-calculated based on an updated dataset.

Among the adaptive approaches listed in Kadlec et al. (2011), *moving windows* methods have been judged as the most suitable for the monitoring set-up used in the study. In this study, two possible approaches are considered: (i) *block-wise* moving window and (ii) *step-size* moving window. In the first case, the parameters of the adaptive methods (i.e. the loadings $p_{i,j}$) are re-estimated at fixed time interval, e.g. on a weekly or monthly bases. For example, it can be hypothesized that DQC of data is carried out by the operator on a weekly basis, and that the re-estimation is subsequently performed by using the newly validated measurements with the same frequency. Conversely, the data sample in *step-size* moving windows is updated every time a new validated measurement is available. Given the high temporal resolution of water quality measurements (usually in the order of 1 minute), this approach can be regarded as computationally expensive, so it is more realistic that the parameter estimation is carried out at lower temporal resolution (e.g. on hourly or daily basis). The step-size approach is more suitable for automatic DQC routines, while the block-wise is more suitable for approaches when the operator interaction is required.

Performance evaluation.
Different aspects were investigated to assess the performance of the sensor:

- *Validity of the multivariate approach*, i.e. is the hypothesis that a combination of on-line water quality measurements can be used to estimate another variable valid? This was assessed in an off-line set-up, where all the observations used for validation were known, i.e. the mean μ_m and the variance σ^2_m were calculated over the entire validation period. The *m-th* variable is projected in the m-dimensional space, and its value is then calculated by using the *m-1* variables as in eq. 4.
- *Size of the adaptation window*, i.e. how long historical data series are necessary to calibrate the sensor? This was assessed by increasing the size N of the data sample used to estimate the loadings $p_{i,j}$).
- *Reliability of predictions over time*, i.e. for how long does the sensor provide reliable predictions? This was evaluated by calculating a performance indicator for validation periods of increasing size. The Mean Absolute Relative Error (MARE – Bennet et al., 2013) was selected to ensure that both dry and wet weather periods (characterized by different magnitude of measured concentrations) were weighted equally.

$$MARE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x^*_{i,m} - obs_{i,m}}{obs_{i,m}}\right| \qquad (5)$$

- *Adaptation methods*, i.e. which moving window approach ensures better performance of the sensor: block-wise or step-wise? To test the first approach, the available dataset was subdivided in 63 weeks (with calibration of sensor used as additional criterion to split the sample, avoiding that the data before and after calibration were store into the same week). To simulate realistic procedures, where data are validated by operators on a weekly basis, loadings and data statistics μ_m and σ^2_m were calculated based every week. In the step-size approach, assuming an automatic data validation procedure which does not require user interventions, the loadings and μ_m and σ^2_m were calculated on a daily basis

RESULTS AND DISCUSSION

Validity of the multivariate approach.

Figure 3 below shows the evolution of MARE after the parameter estimation for adaptation windows of different sizes. The left figure illustrates how the median MARE was below 10% until 6-7 weeks after the parameter estimation, with a clear degradation of the performance of the sensor over time. Interestingly, better performances were obtained with shorter adaptation windows (2-3 weeks), whereas the longest windows (8 weeks) showing a median MARE above 15% after 10 weeks from parameter estimation. The same value was exceeded by the 2-weeks adaptation window only after 14-15 weeks. This result is likely to be caused by the quality of the available data: the changes in the measured concentrations due to sensor degradation and subsequent sensor re-calibration (such those shown in Figure 1) affect the performance of the multivariate software sensor. Wider adaptation windows will include data collected during periods which are no longer representative of the current sensor readings, leading to worse performance. Conversely, shorter adaptation windows consider only the most recent measurements, thus providing better estimation of the missing values.

It is important to stress that these considerations are based on median values: the short adaptation window might not include sufficient information to predict the behaviour of the measured variable for specific events (for example, if the adaptation window contains only data from dry weather period, the software sensor could encounter difficulties during a rain event). This is shown in the Figure 3 (right): the shorter adaptation windows (2-4 weeks) showed a wider spread of MARE (with cases of MARE up to about 100%), while the wider adaptation windows showed more consistent performance (i.e. a narrower spread of the results).

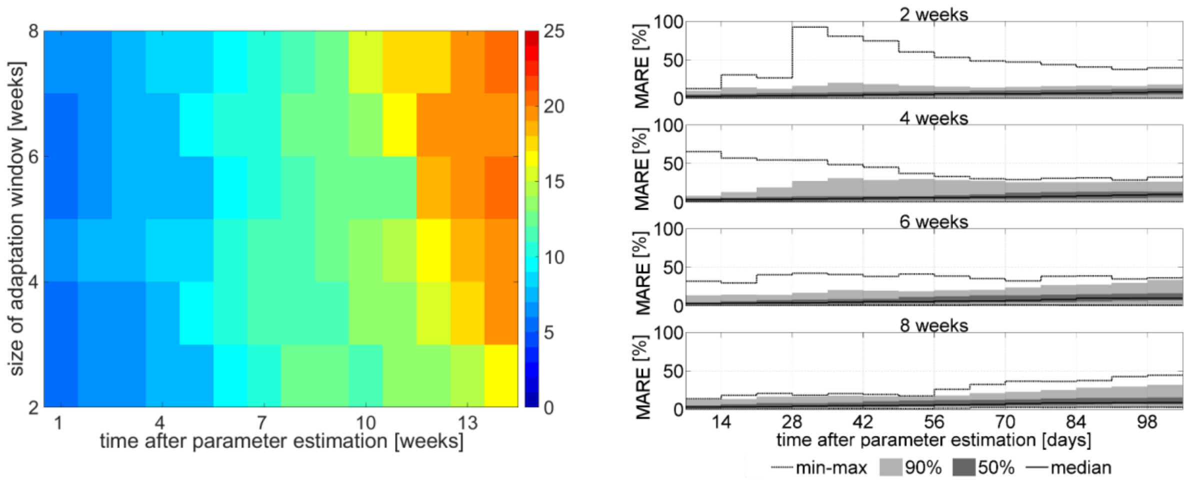

Figure 3. Statistics of software sensor for NH_4 in an off-line set-up (μ_m and σ^2_m known). Left: value of MARE [%] (median of values estimated over the entire dataset) over time for adaptation windows of different size. Right: Temporal evolution of MARE (value estimated on a weekly basis) after parameter estimation for adaptation windows of different size.

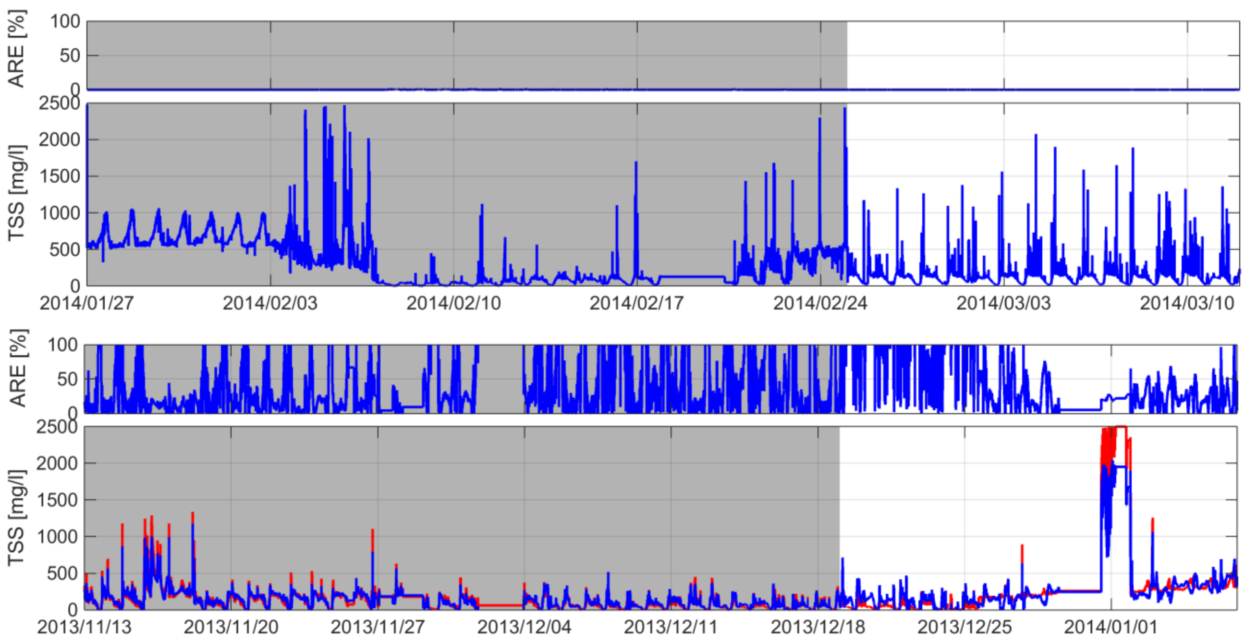

Figure 4. Examples of measured (red) and simulated (blue) TSS concentrations for selected periods, along with the calculated Absolute Relative Error. The adaptation window is shown in grey. Above: the period with the best MARE value. Below: period with the worse performance in the dataset.

The results shown in Figure 3 refer to NH_4, which is a dissolved component of the wastewater flow, i.e. it can be easily be predicted by using a simple dilution approach (see the examples in Langeveld et al., 2014 and Sharma et al., 2014). The results obtained for TSS showed also good performances: the MARE value were generally 10% higher than those obtained for NH_4, but the predicted TSS concentrations are still acceptable. Figure 4 shows the periods with the best and the worst MARE: overall, the multivariate software sensor approach seems to be suitable to predict water quality variables based on a combination of other measurements, i.e. the information carried out by four water quality variables is sufficient to calculate a fifth variable with an acceptable error. This result suggests that the multivariate software sensor could be used for on-line data validation, extending the application of the multivariate approach presented in Alferes et al. (2013), which was applied off-line, to on-line settings.

Application in on-line conditions

Reliability over time. The use of the software sensor in realistic conditions, when the measurement is not available and it is interpolated (step 5 in Figure 2), showed an important decrease in the performance compared to the off-line case (where the measurement was reconstructed from its projection). In the case of block-wise (Figure 5), the MARE medians for ammonia were above 40% after 4 weeks since parameter estimation. Similarly to the off-line case, the best median values were obtained with shorter adaptation window, with wider windows leading to a more rapid deterioration of the software performance. The TSS sensor showed worse performance than the ammonia sensro, with median MARE never below 100%.

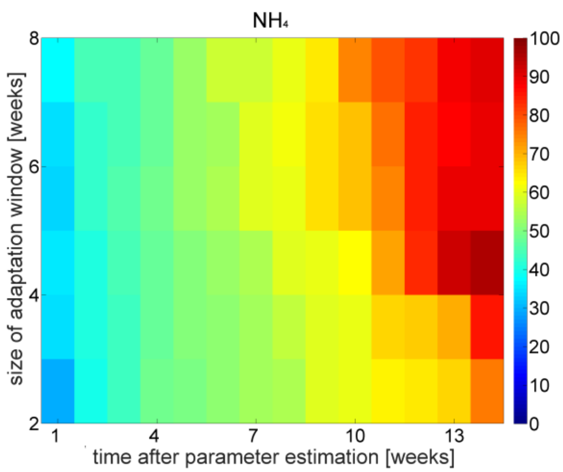

Figure 5. Evolution of median MARE [%] for block-wise adaptation window of different sizes for NH_4 (left) and TSS (right). Please note the different scales.

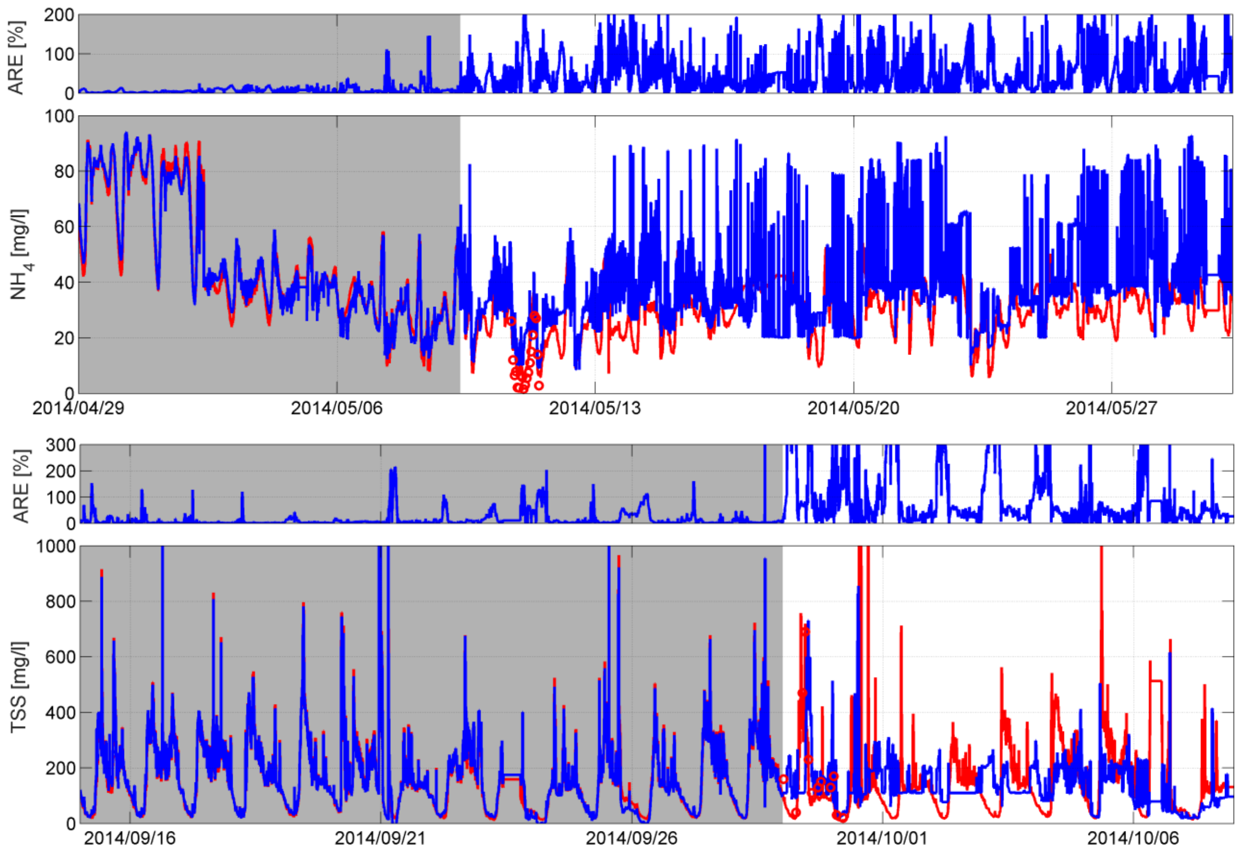

Figure 6. Examples of measured (red) and simulated (blue) NH_4 and TSS concentrations, along with Absolute Relative Error [%]. Measurements collected by autosampler are shown by red circles. The adaptation block-window is shown in grey.

The rapid degradations of the software sensor are exemplified in Figure 6: for both ammonia and TSS the error significantly increases compared to the calibration period, exceeding 100% about 5-7 days after the parameter estimation for NH_4 and 2-3 days for TSS. This suggest a poor predictive ability for this approach for the particulate pollutant. Overall, it can be concluded that both software sensors are not capable of providing reliable results for periods of time greater than one week after the parameter estimation.

Adaptation method. When comparing the block-wise and the step-wise approaches, their performance over the entire dataset is comparable. Figure 7 shows the comparison for the MARE estimated on the first day after parameter estimation: for ammonia, the step-wise approach provides slightly better performance when looking at the median value (24.4% and 20.7%, respectively) and the best MARE (7.52% and 1.64%, respectively). The degradation of the sensor performance for the block-wise approach is shown in Figure 8, where it can be seen that the daily MARE does not increase linearly, but it stabilizes 2-3 days after the parameter estimation. The poor performance of the TSS sensor for both the approaches confirms that the proposed software sensor is not suitable for on-line prediction of pollutants not strictly linked to wastewater sources. These results suggest that, whenever possible (i.e. when on-line automatic data validation is available), the step-wise adaptation method should be preferred.

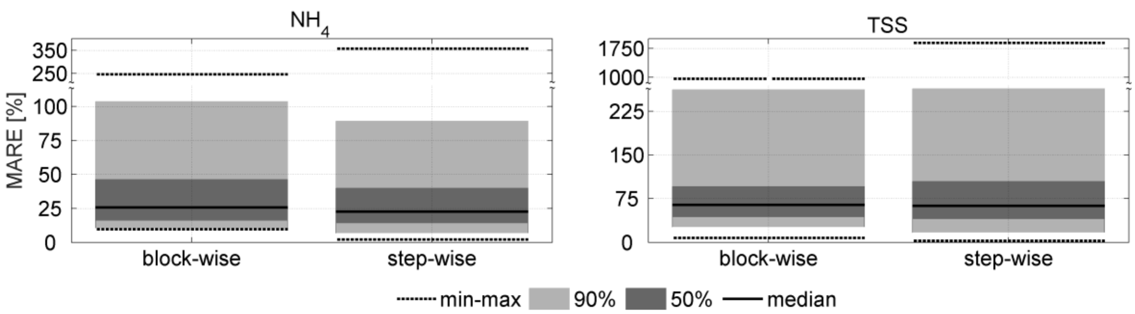

Figure 7. Value of MARE estimated for the first day with different adaptation methods. Note the different scale on the y-axis for NH_4 (left) and TSS (right).

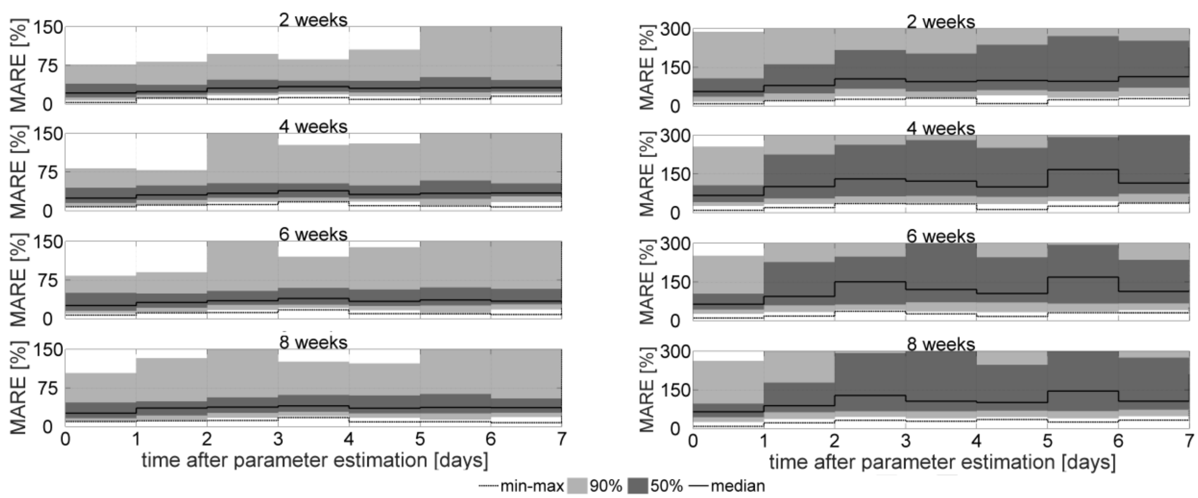

Figure 8. Temporal evolution of MARE (daily value) in the period between two parameter estimations for block-wise adaptation. Note the different scale on the y-axis for NH_4 (left) and TSS (right).

Future outlook

The PCA-based software sensor presented in this study is a first attempt to apply in on-line conditions techniques which are widely applied in an off-line context. A number of limitations of the presented approach can be identified, which suggest possibilities for further investigation and development.

A major factor affecting the results of this study is the quality of the dataset collected during the AMOK project: the main assumption behind the software sensor is that the data in the adaptation window are representative of the current situation. This assumption is clearly undermined when sudden changes in the sensor readings are introduced in the dataset after calibrations (as shown in Figure 1). Furthermore, short adaptation windows may explain the poor performance in the prediction of TSS concentrations, especially during wet weather periods. When the adaptation window includes only a limited number of rain events, the software sensor has not sufficient knowledge of the interaction between TSS and the other measured variables, leading to poor results in prediction. A possible solution to this problem may be provided by a *piecewise* formulation of the adaptation window: the dataset used to estimate the sensor parameters should be subdivided into dry- and wet-period, with the data contained in the latter updated only when a new rain events is recorded.

The application of the software sensor to a dataset of better quality might lead to better results. On the other hand, everyday operations of on-line sensor in sewer systems do not ensure the same quality standards of extensive monitoring campaigns (more accurate, but also requiring greater resources), and they are thus likely to provide data of similar quality to those used in this study. The performance estimated in this study thus provides an overview on the potential of the software sensor in real-life conditions.

Similarly, only simple approaches were used to validate the timeseries, leaving an evident noise in the measurement (e.g. for turbidity measurements), which could affect the performance of the software sensor. These may be improved by the application of advanced DQC approaches, such as the univariate method described in Alferes et al. (2013). However, it should be stressed that the DQC methods should be able to run on-line and with minimum requirement for user interaction.

The PCA-interpolation method used in this study is quite basic. A wide range of more complex methods are available in literature to estimate missing measurement through PCA (Ilin and Raiko, 2010). These methods, ranging from iterative to probabilistic approaches may significantly improve the predictive capability of the multivariate software sensor. However, the need for an on-line application may pose a barrier for the use of the most computationally expensive approaches.

CONCLUSIONS

This study evaluated the performance of multivariate software sensors (based an adaptive PCA) when applied for on-line data reconstruction of sewage quality measurements. The results obtained after the performance evaluation which was carried out on a period of about 15 months suggest that:

- Adaptive PCA methods can provide a good estimation of water quality variables over long periods of time (10 weeks) when measurements of all the variables are available. This shows how PCA-based DQC methods, which are widely applied in an off-line context, can successfully be applied on-line.
- The predictive power of the software sensor was quite limited in time (1 week) when used to reconstruct missing measurements.
- Methods using a step-wise adaptation window, where the software sensor parameters are re-estimated on a daily basis, provide better performances than methods using a block-wise adaptation window, i.e. where the parameters are updated at a lower frequency.

Overall, this study highlights the potential for the development of on-line multivariate software sensors. These can improve the quality of the water quality measurements collected in sewer systems, thus boosting the application of these sensors for on-line optimization of urban drainage systems.

ACKNOWLEDGEMENT

This study was supported by the project AMOK-Advanced Monitoring of Sewage Quality, which was financed by the Foundation for Development of Technology in the Danish Water Sector (VTU-Fonden). Part of the software sensors were developed during Luca Vezzaro's Industrial Postdoc project "*Model-predictive control of urban water systems under uncertainty*", which was financed by the Innovation Fund Denmark. The authors would like to thank dr. J.B. Aubin (INSA-Lyon) for his contribution to the discussion on the applicability of PCA methods.

REFERENCES

Alferes, J., Tik, S., Copp, J., Vanrolleghem, P.A. (2013). Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. *Water Sci. Technol.*, **68**(5), 1022-1030.

Bennett,N.D., Croke,B.F.W., Guariso,G., Guillaume,J.H.A., Hamilton,S.H., Jakeman,A.J., Marsili-Libelli,S., Newham,L.T.H., Norton,J.P., Perrin,C., Pierce,S.A., Robson,B., Seppelt,R., Voinov,A.A., Fath,B.D., and Andreassian,V. (2013) Characterising performance of environmental models. *Env. Model. Softw.* 40 1-20.

Campisano, A., Cabot Ple, J., Muschalla, D., Pleau, M., Vanrolleghem, P.A. (2013). Potential and limitations of modern equipment for real time control of urban wastewater systems. *Urban Water J.*, **10**(5), 1-12.

Gruber, G., Winkler, S., Pressl, A. (2005). Continuous monitoring in sewer networks an approach for quantification of pollution loads from CSOs into surface water bodies. *Water Sci. Technol.*, **52**(12), 215–223.

Kadlec, P., Gabrys, B., Strandt, S. (2009). Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng*. **33**(4), 795-814.

Kadlec, P., Grbic, R., Gabrys, B., (2011). Review of adaptation mechanisms for data-driven soft sensors. *Comput. Chem. Eng*. **35**(1), 1-24.

Ilin, A., Tapani, R. (2010). Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *J. Mach. Learn. Res.*, **11**, 1957–2000.

Lacour, C., Schuetze, M. (2011). Real-time control of sewer systems using turbidity measurements. *Water Sci.Technol.*, **63**(11), 2628-2632.

Langeveld, J., Schilperoort, R., Van Daal, P., Benedetti, L., Amerlinck, Y., De Jonge, J., Flameling, T., Nopens, I., Weijers, S. (2014). A new empirical sewer water quality model for the prediction of WWTP influent quality. Proceedings of the 13th International Conference on Urban Drainage (13ICUD), Kuching, Sarawak, Malaysia, 7th-12th September 2014

Li,J., Heap,A.D. (2014) Spatial interpolation methods applied in the environmental sciences: A review. Env. Model. Softw. **53**, 173-189.

Metadier, M., Bertrand-Krajewski, J.L. (2012). The use of long-term on-line turbidity measurements for the calculation of urban stormwater pollutant concentrations, loads, pollutographs and intra-event fluxes. *Water Res*. **46**(20), 6836-6856.

Rosen, C., Lennox, J.A. (2001). Multivariate and multiscale monitoring of wastewater treatment operation. Water Res., **35**(14), 3402-3410.

Sharma, A.K., Lynggaard-Jensen, A., Vezzaro, L., Andersen, S.T., Brodersen, E., Eisum, N.H., Snediker Jacobsen, B, Høgh, J., Gadegaard, T.N., Mikkelsen, P.S., Rasmussen, M.R. (2014), Advanced Monitoring of Combined Sewer Overflows: what to measure and how to measure. Proceedings of the 13th International Conference on Urban Drainage (13ICUD), Kuching, Sarawak, Malaysia, 7th-12th September 2014